Improving Metacognition in a College Classroom: Does Enough Practice Work?

Christopher A. Was
Kent State University

Tara L. R. Beziat
University of South Carolina Aiken

Randy M. Isaacson
Indiana University South Bend

Abstract

Metacognition is often described as knowledge and control over one's cognitive processes. Models of metacognition often include knowledge monitoring as the foundation of metacognitive skills. The current study was designed to determine whether the ability to accurately assess one's knowledge can increase throughout a semester long course, when students are provided knowledge monitoring practice. Undergraduates' enrolled in an educational psychology course were administered 13 exams during the course of a semester and provided a number of opportunities to practice knowledge monitoring. Prior to each exam students were required to predict their exam scores. Calibration (the difference between predicted scores and actual performance) improved over the course of the semester. However, the data also revealed improved calibration might have been an artifact of the data. Put differently, calibration was poor at the beginning of the semester as students were on average overconfident. By the end of the semester, students predicted scores had not changed, but exam scores increased thus improving calibration.

Knowledge monitoring is a basic metacognitive process essential to learning. Imagine a student preparing for an upcoming examination in her educational psychology course. To prepare efficiently and to be well prepared for the exam, the student must be able to identify those concepts she has already mastered and those concepts that will require more effort and study time. This ability to monitor one's own knowledge is a key to metacognitive and self-regulation processes during learning. Indeed, Tobias and Everson (2009) proposed a hierarchy of metacognitive processes with monitoring knowledge as the foundation. In the Tobias and Everson (2009) model, higher-level metacognitive processes, such as selecting strategies, evaluating learning, and planning are dependent upon accurate knowledge monitoring. Tobias and colleagues have demonstrated that the

ability to accurately judge one's knowledge (knowledge monitoring accuracy) is predictive of math achievement, reading achievement and even GPA scores (see Tobias & Everson, 2009 for a review).

Other theories of metacognition also propose that effective knowledge monitoring leads to better regulation during studying (e.g., Metcalfe, 2009; Nelson & Narens, 1990). One goal of the current investigation was to determine if individual differences in knowledge monitoring accuracy are related to academic success within a classroom setting. A second goal was to determine if training improves students' knowledge monitoring accuracy as measured by calibration.

The Knowledge Monitoring Assessment

As noted by Serra and Metcalfe (2009) metacognition is not flawless and poor metacognition can have a negative impact on studying and performance. Previous research has demonstrated that college students who are better at knowledge monitoring, as measured by predicting scores on an exam (i.e., calibration) are also likely to outperform those students who are not accurate knowledge monitors (Hacker, Bol, Hogan & Rakow, 2000; Isaacson & Fujita, 2001). Hartwig, Was, Isaacson, & Dunlosky, (2012) demonstrated a clear connection between knowledge monitoring accuracy and academic performance. In their investigation Hartwig, et al. (2012) developed a knowledge monitoring assessment based on the method presented by Tobias and Everson (2002; 2009).

The knowledge monitoring assessment used by Hartwig, et al. (2012) required participants to judge whether they knew the definition of a word or not. Participants made a yes (known) or no (not known) judgment for each of 50 vocabulary words, and were then required to complete a multiple choice test in which they were presented with each of the 50 vocabulary items and five possible synonyms. Four of the possible synonyms were distractors and a fifth was an actual synonym of the vocabulary items. The knowledge monitoring assessment generated the following possible outcomes. Students indicate the word is: 1) known and correctly responded to the item on the vocabulary test [hits]; 2) known but responded to incorrectly on the test [false alarms]; 3) unknown but the correct response was given on the test [misses]; and 4) unknown and responded to incorrectly on the test [correct rejections]. Hits and correct rejections represent accurate knowledge assessment, whereas false alarms and misses represent inaccurate knowledge assessments.

Hartwig et al. (2012) administered this knowledge monitoring assessment in the first two weeks of the semester of an undergraduate course in educational psychology and found that accuracy on the knowledge monitoring assessment was correlated to final exams

scores. Although the correlation between the knowledge monitoring assessment and final exam score was moderate (r = .39) it represented a substantial amount of variance in final exam scores when one considers the knowledge monitoring assessment was completed at the start of the semester and the final exam was administered at the end of the semester. Furthermore, the number of variables that might influence final exam performance is quite large. The finding that the knowledge monitoring assessment accounted for variance in final exam scores is therefore notable. Indeed, Hartwig et al. (2012) split the participants into quartiles based on the knowledge monitoring assessment scores and found the quartiles differed in exam performance such that students who monitored more accurately also earned higher grades, on average, on the final exam.

Improving Monitoring Accuracy

The results of Hartwig, et al. (2012) and others (e.g., Hacker, et al., 2000; Isaacson & Fujita, 2001) provide evidence that knowledge monitoring accuracy is related to performance on exams. Although these findings are important, it is even more important to know if knowledge monitoring accuracy can be improved. If successful knowledge monitoring leads to positive academic outcomes, it follows that teaching students to be better knowledge monitors would make them better and more successful students. In an attempt to determine if students' knowledge monitoring accuracy could be improved through pedagogical practices, Isaacson and Was (2010a) designed a classroom study in which they measured the knowledge monitoring accuracy at the beginning and the end of the semester of 106 undergraduates enrolled in an educational psychology class. Throughout the semester the students were required to frequently make monitoring judgments about their knowledge. Several opportunities were provided to the students to practice knowledge monitoring (cf. Isaacson & Was, 2010b). The most important of which was a weekly variable-weight and variable-difficulty exam (the exam format is described in detail in the methods section).

Isaacson and Was (2010a) used the same stimuli in both administrations of the knowledge monitoring assessment. It was found that the knowledge monitoring assessment completed at the beginning of the semester and the one completed at the end of the semester were both correlated to the score on the final exam in the course. This finding supports the conclusions of Hartwig, et al (2012). More importantly, Isaacson and Was (2010a) found a significant increase in students' knowledge monitoring accuracy from the knowledge monitoring assessment scores at the beginning of the semester to scores at the end of the semester. Isaacson and Was (2010a) proposed that the weekly monitoring practice provided throughout the semester increased students' general knowledge monitoring ability.

In an attempt to replicate the findings of increased knowledge monitoring accuracy over the course of the semester, Was, Issacson, Beziat, and Dippel (2011) conducted a study using the same methodology. Again, a significant increase in knowledge monitoring accuracy was found. However, Was et al. (2011) discovered that although there was a significant increase in the number of hits and a significant decrease in the number of misses, the rate of false alarms did not change. Therefore, the increase in gamma may reflect an artifact of the data. Put differently, if students are overconfident in their knowledge assessments the increase in hits at the end of the semester may reflect an increase in knowledge (i.e. items answered correctly), not an increase in accurately identifying known items. This may indicate that students have difficulty changing an optimistic bias or overconfidence (Hacker et al., 2000). The lack of change in the rate of false alarms and the increase in hits raised two important questions.

Overconfidence

The first question is whether there is a general overconfidence bias in students' knowledge monitoring? The most common method in the extant literature used to measure knowledge monitoring within a classroom context is calibration between exam score prediction and exam scores (e.g., Hacker, Bol, Hogan & Rakow 2000; Isaacson & Fujita 2001, Miller & Geraci, 2011). Calibration is operationalized as the difference between predicted performance and actual performance. A common, yet not surprising finding involving undergraduate students, is a striking difference between high and low performing students in their ability to predict their test scores. Typically, successful students demonstrate better calibration, whereas poorer performing students over-estimate their future performance. For example, Hacker et al. (2000) administered three multiple-choice exams to undergraduates over the course of a semester. Before each exam, students were required to predict their test scores. Immediately following the exam, but before it was graded, students again estimated their test scores (postdiction). Results indicated that the highest performing students were more accurate in their predictions of exam scores as well as the post-diction of performance. In turn, the lowest performing students' calibration was poor in both prediction and postdiction of exam scores, with the lower performing students greatly over estimating their performance even after completing the exam.

Isaacson and Fujita (2001) administered 10 weekly examinations to undergraduate students over the course of a semester. Again, lower achieving students had a tendency to make predictions that were higher than their actual test scores. Overconfidence bias was also demonstrated in an investigation conducted by Vadhan and Stander (1994). The results of these studies suggest that high performing students are able to predict how they are going to do on a test and can also accurately assess how they have performed.

However, in general, students have a tendency to be overconfident when predicting their test scores, with the lower performing students having the most difficulty with calibration, with a tendency to be overconfident. Clearly, knowledge monitoring accuracy, as measure by calibration, has an impact on students' academic outcomes.

Improvement

This leads to our second question. Can classroom practices decrease students' overconfidence? There is inconsistency in the literature regarding the improvement in students' ability to predict their performance on test of knowledge and understanding. For example, Hacker, et al. (2000) found that undergraduates' predictions of exam scores were more accurate on a third exam as compared to the first exam. However, the third exam was a cumulative examination of material contained in the first and second exam, and this may in part account for the increased accuracy. Contrary to the Hacker et. al. (2000) results Bol, Hacker, O'Shea and Allen (2005), also, Nietfeld, Cao and Osbourne (2005) found no improvement in monitoring accuracy even after a semester of monitoring practice, but a more recent study conducted Nietfeld, Cao and Osbourne (2006) found that an intervention of monitoring exercises and feedback had a significant impact of students' calibration and test performance.

In a recent investigation involving undergraduate in two semester long studies, Miller and Geraci (2011) again found that students were overconfident in their predictions of test scores and again the lower performing students were particularly poor at predicting their test scores. Germane to the current study, Miller and Geraci (2011) attempted to increase metacognition (as measured by improved calibration) by providing incentives for calibration accuracy and feedback regarding how to improve calibration. The data from Experiment 1 indicated that providing incentives and only minimal feedback did not improve calibration or exam performance. However, in Experiment 2 increasing the salience of the feedback increased calibration for lower performing students without increasing their exam performance.

The investigation conducted by Miller and Geraci (2011) had two limitations that may have contributed to their limited findings. First, in both experiments, Miller and Geraci administered only four exams across the semester. This provided limited opportunities for the participants to practice predicting their test scores.

Second, Miller and Geraci (2011) required students to record a letter grade as the prediction of their exam outcomes (e.g. "A-"). For analyses this letter grade prediction was converted into a numeric value based on the grading scale used in the course. For example, if a student recorded a "B+" that prediction would be converted into an 88%. A prediction of "B" was converted to 85% as that was the midrange of a B on the grading scale. To calculate calibration, the percent correct on the exam was subtracted from the

converted prediction and divided by 100. This was then subtracted from one and multiplied by 100 to account for the fact that 100% was the maximum percentage correct. The students participating in the two experiments were informed via the course syllabi that they could earn two percentage points extra credit for each of the four exams if they predicted any version of the grade earned. For example, if a student predicted an "A" but received an "A-" they would be given the extra credit. The formula used to measure calibration and the awarding of credit for limited accuracy may have contributed to the lack of substantial improvement in both calibration and performance. For example, the student who predicted a B+ (88%) but received a B- (82%) would receive the credit, but the student who predicted a C+ (78%) and received a B- (82%) would not. Thus the less accurate student in this case would receive positive feedback and reinforcement for being less accurate.

Another contributor to the lack of change in calibration in the Miller and Geraci (2011) investigation may have been the treatment used to improve calibration. The instruction given to participants to improve their predictions was that improving their scores (performance) or lowering their predictions would improve calibration. It is unlikely that such feedback would increase actual metacognition. Although the lower performing students did state that they increased their studying or lowered their predictions, this does not translate to better understanding of knowledge monitoring or metacognition. The most common response among high performing students was that the feedback did not influence their predictions.

Goal of the Current Investigation

The current investigation was undertaken in order to improve upon what we see as limitations in the extant literature. To date, researchers' attempts to determine if practice could improve calibration have provided students limited opportunities to practice predicting the outcomes of examinations, limited and delayed feedback on performance, and a focus on improving predictions, not improving metacognition. In the current investigation, we provided students with much more opportunities to practice knowledge monitoring and reflect on their own knowledge than any study we were able to find. Furthermore, we feel that extensive practice and training is necessary to increase students' metacognition, beyond simply decreasing the difference between predicted test scores and actual performance.

We conducted the current investigation to determine if more practice would lead to improved metacognition as measured by calibration. It was our hypothesis that weekly practice of prediction and postdiction of test scores, and the opportunity to reflect on calibration based on immediate feedback, would improve students' calibration.

Methods

Participants:

250 students enrolled in and introductory educational psychology course participated in exchange for course credit. Females represent 77% of the participants. All students did not complete every exam and/or every prediction questionnaire and therefore, there is missing data. All analyses were completed using listwise deletion.

Design and Procedure:

Weekly Examinations: Students were administered weekly objective examinations throughout the duration of the semester in which they were enrolled in the course for a total of 13 examinations. Each examination was based on a variable weight, variable difficulty format. Each examination contained a total of 35 questions composed of 15 Level I questions that were at the knowledge level, 15 Level II questions at the evaluation level, and 5 Level III questions at the application/synthesis level. Scoring of the exam was based on a system that increased points for correct responses in relation to the increasing difficulty of the questions: Level I questions were worth 2 points each, Level II questions were worth 5 points each, and 5 Level III questions were worth 6 points each. Students were also required to choose the questions they were least confident about and these questions were only worth one point (5 of the 15 Level I and II questions, and 2 of the 5 Level III questions). The scoring equaled a possible 100 points for each exam. Correlations between total score and absolute score (number correct out of 35) ranged from $r = .87$ to $r = .94$. Therefore, all analyses were completed using total score.

Knowledge Monitoring Practice Opportunities.

Throughout the semester long course, students were presented with a number of resources in the curriculum to improve knowledge monitoring. For example, students were encouraged to take on-line practice quizzes each week that have a format similar to the weekly exams (variable weight and variable difficulty) in which students are asked about their confidence of each answer before the practice quiz was graded on-line. The course also used a web-based course management system with a variety of resources developed to improve metacognition (e.g., students completed weekly self-reflections which focused on self-regulated learning and metacognition). The course had small discussion classes led by peer mentors where students were given a quiz each week also using a format similar to the weekly exams. Students also submitted a journal to their peer mentor each week that focused on self-regulated learning and metacognition. The class had two lectures each week and students were presented with a Question of the Day at the start of every class with their answer to these questions recorded using a student response system (i.e., ”clickers”) that required students to indicate whether they are absolutely sure, fairly sure, or just guessing at the answers. Students could earn 200 points (8% of the total course grade) across the semester for their Question of the Day

responses. Students earned points for correct answers, but also for accurate knowledge monitoring. For example, if a student indicated she was absolutely sure, she earned 9 points if she was correct, but no points if she was wrong. However, if a student indicated he was unsure or just guessing, he earned 3 points if he was correct and 2 points if he was wrong.

Calibration:

Prior to beginning each exam students completed a pre-test questionnaire asking them to predict the total number of points they would receive on the exam. Immediately following the examination students completed the remainder of the questionnaire requiring them to indicate the total number of points they believed they had earned. Exams were then immediately scored for the student using an Apperson® test scoring scanner. Students were then allowed to review their exams, predictions as postdictions.

As an incentive to increase calibration accuracy, students were awarded two extra points toward their exam score if the accurately predicted their test scores, and two points if they accurately postdicted their exam score. One point was awarded for both prediction and postdiction if students were within one point of their score (e.g. if a student predicted a 90, she would receive one extra point if her exam score was between 89 and 91).

Results

Table 1 presents the means and standard deviations of exam scores, predicted scores, and calibration across the semester. Two participants were removed from the analyses. The first was removed because the participants' mean calibration score across the semester was greater than four standard deviations above the mean. The second participant was removed because his or her mean calibration was four standard deviations below the mean. The current analysis is based on 12 of the 13 exams completed during the semester because the last weekly exam was not included in the analysis. The course syllabi allowed for students to drop one exam score and the majority of the students (over 65%) choose not to take Exam 13 and the mean score of those who did was far below the mean of the other weekly exams. Reliability analysis revealed that total scores were reliable across the 12 included exams ($\alpha = .93$) as were predicted scores ($\alpha = .96$).

Calibration was measured as the difference between the predicted test score for each exam and actual total points earned on that respective exam. Therefore, a positive calibration score represents an overestimate of performance and a negative calibration score represents an underestimate of exam performance. A calibration score of zero reflects perfect calibration of prediction and test performance. We chose this simple calibration score due to ease of interpretation. For example, a calibration score of 6 indicates that the student predicted she would get a score six points higher than the actual score obtained on the exam. This represents overconfidence. We created a mean

calibration score by averaging each students calibration scores for all exams (*M* =2.04) and determined that on average students were overconfident, $t(248) = 5.61$, $p <.001$, Mean Difference = 2.04, CI = 1.33; 2.75. Calibration and exam scores were averaged across exams and were found to have a strong correlation, $r = -.62$, $p < .001$. This negative correlation indicates that as calibration scores decrease exam scores increase. The scatterplot presented in Figure 1 graphically represents this relationship. The line positioned at 0 on the Y-axis indicates perfect calibration.

To further examine this relationship we calculated the mean of all calibration scores and the mean of all exam scores across the semester for each participant. We then divided participants into two groups based on the mean calibration score (*M* = 2.04). Figure 2 presents the mean examination scores across the semester for the mean groups. We conducted an independent samples *t*-test to determine if the mean exam scores across the semester were different for those above and below the mean calibration score. The mean exam score for those above and below the mean calibration score were *M* = 86.06 (*N* = 134) and *M* = 76.59 (*N* =127) respectively. The Levene's test for equality of variance revealed inequality of variance, $F = 39.10$, $p < .001$. We therefore report the *t* value with equality of variances not assumed. This analysis revealed a significant difference in average test scores between the calibration mean-split groups, $t(190.07) = 9.01$, $p < .001$(CI: 7.39, 11.54). There is a clear difference in test scores across the semester for students above and below the calibration mean with students scoring above the mean scoring lower on exams on average than students below the calibration mean. Put differently, students who were more accurate, or even under predicted their test scores performed better than those less accurate and overconfident in their predictions.

As in previous studies, our data demonstrate that students performing at the highest levels more accurately predict their future performance, with a tendency to underestimate, whereas the poorest performing students are poor calibrators with a tendency to overestimate future performance with a greater magnitude of error. This is also the case for each exam measured separately (Appendix A).

The major focus of the current study was to test the hypothesis that extensive practice at calibration would increase students' ability to accurately predict performance. Figure 3 displays mean calibration score on each exam. The line positioned at zero on the Y-axis represents perfect calibration. As is evident from the figure, students' calibration accuracy improved as the semester progressed. Table 2 displays a series of *t*-tests completed to determine if calibration scores were significantly different from zero. Analyses indicated that calibration scores on exams 1 - 8 were significantly different than 0. Of these, all calibration means were above zero (indicating students were overconfident) with the exception of Exam 6. The mean calibration score for Exam 6 was below zero. Most importantly, for Exams 9, 11 and 12 the mean calibration score did not differ from zero. The mean calibration score at Exam 10 was significantly different than

0, but the calibration mean was below 0 and not above. These results, although based on a null result, indicate that by exam 9 students on average were accurate predictors of exam performance.

Discussion

The current data support the conclusion of previous research that students performing poorly on objective examinations are likely to overestimate their performance. The current data also align with previous findings that indicate the best performing students are more accurate in their predictions of performance, if not slightly under confident. More important, the data provide evidence that practice may support the development of effective metacognitive knowledge monitoring. Put differently, calibration is a metacognitive skill and students provided with regular practice can improve this skill. Perhaps opportunities to practice calibration (e.g., quizzes, exams, self-testing and reflection) in turn influence higher order metacognition and self-regulation.

The current results also expand upon the extant literature. Previous findings regarding increased knowledge monitoring accuracy were minimally impressive at best and were often based on limited metacognitive practice. In the current study, we provided students with what we consider deliberate metacognitive practice. We did not simply have students practice predicting their test scores (although that was a part of our procedures), but we also had them practice simple knowledge monitoring strategies. For example, having students regularly judge whether they were absolute sure, somewhat sure, or just guessing in response to each question. We also had students practice more deep processing of their metacognition. The weekly journals in which students wrote about their metacognitive and self-regulating strategies were designed to encourage this type of deep processing.

A further contribution to existing literature was made by using calibration between predicted test score and actual test score as the measure of increased knowledge monitoring. Many previous investigations as noted above had used this measure with a minimal number of tests (e.g, Hacker, et al., 2000; Miller & Geraci, 2011). Others had used similar extensive, deliberate metacognitive practice, but used an external measure of knowledge monitoring (e.g, Hartwig, et al., 2012; Isaacson & Was, 2010a) not calibration of exam scores. We feel these are important contributions to the understanding of knowledge monitoring as a trainable skill.

Although the results of our investigation are encouraging, they must be interpreted with caution. Figure 4 displays the mean predicted test scores and the mean actual scores across the semester. Review of Figure 4 suggests that the increase in calibration is not a result of better knowledge monitoring, but instead a result from students' test scores increasing. Put differently, as is evident in Figure 4, test scores changed dramatically over

the semester. The mean exam score across exams was 82.94 with a standard deviation of 2.95. The lowest mean test score occurred at test 1 and was 76.52. The mean score of the last exam of the semester was 84.22.[1] This is stark contrast to the predicted scores of which the mean across the semester was 84.43 with a standard deviation of .71. The lowest mean predicted score was 82.66 and the highest was 85.53. The change in calibration may be simply the increase in mean test scores over the course of the semester, whereas the predicted scores did not change. However, another interpretation is that increased knowledge monitoring lead to an increase in test scores. As instructors, we were pleased to see this increase in test scores. However, as investigators we were disappointed that we did not have conclusive evidence of an improvement in knowledge monitoring.

Implications

The results of this study support the idea that providing multiple opportunities for metacognitive practice leads to better knowledge monitoring. Based on these results, it is possible for classroom teachers to improve their students' knowledge monitoring and in turn their academic performance. In order to do this the classroom teacher must provide a significant number of opportunities for the student to practice their knowledge monitoring and the student must receive prompt and informative feedback about their performance.

Evidence from the current research also suggests that poor performing students can improve their knowledge monitoring when provided ample practice. As previous research has shown poor performing students often overestimate their performance on quizzes and exams (Isaacson & Fujita, 2001; Vadhan & Stander, 1994). The current research provides evidence that when provided with multiple opportunities to practice their knowledge monitoring these poor performing students can improve their calibration and therefore better estimate their performance. If poor performing student continue to improve the accuracy of their knowledge monitoring, this may in turn lead to better preparation for upcoming assessments. When students are more accurate at identifying what they know and what they do not know they tend to perform better on assessments. Taken together, it is possible for teachers to improve the academic achievement of their poor performing students by providing training in knowledge monitoring.

Suggestions for future research

Recall that Isaacson and Was (2010a) and Was, Isaacson, Beziat, and Dippel (2011) found improvement in general knowledge monitoring using a simple knowledge monitoring assessment. However, as with the majority of research interested in improved metacognition, these studies used a measure of relative accuracy (gamma) to measure

change in knowledge monitoring across the semester. Indeed, a great deal of research in metacognition has focused on the accuracy of monitoring through calibration and relative accuracy (Serra & Metcalfe, 2009). Investigations such as that conducted by Miller and Geraci (2011) and the study described here have relied on measures of calibration (the prediction of test scores).

To our knowledge, absolute accuracy of knowledge monitoring as measured by item-by-item confidence ratings, has not been investigated relative to improvement in metacognition and classroom performance. It is evident that a student’s overall sense that she understands the material to be presented on a test would relate to performance on that test. However, as students study and prepare for exams it is likely that they make judgments of learning (JOL’s) on a more item specific basis. Put differently, students may make general JOL’s (e.g., at the chapter level) but are also likely to make more fine-grain JOL’s (e.g., at the definition or concept level). More than one model has been proposed that explains how JOL’s at the item-specific level influence study time and effort (e.g., Dunloksy & Theide, 1998; Metcalfe, 2002). However, there is a lack of research in classroom settings that has examined how these item-by-item judgments relate to performance. We suggest that future research investigate absolute accuracy on exams as a way to capture knowledge monitoring and knowledge monitoring improvement.

References

Bol, L. Hacker, D. J., O’Shea, P., & Allen, D. 2005. The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269-290. doi:10.3200/JEXE.73.4.269-290

Dunlosky, J., & Theide, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica 98*(1), 37-56. doi: 10.1016/S0001-6918(97)00051-6.

Hacker, D., Bol, L., Horgan, D., & Rakow, E. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(4), 160-170.

Hartwig, M., Was, C. A., Isaacson, R. M., & Dunlosky, J. (2012). General knowledge monitoring as a predictor of in-class exam performance. *British Journal of Educational Psychology, 82*(3), 456-468. doi: 10.1111/j.2044-8279.2011.02038.x

Isaacson, R., & Fujita, F. (2001, April). *The effects of goals, expectations, and self-efficacy on self-regulation and performance in college students.* Presented at the Annual conference of the American Educational Research Association, Seattle, Washington.

Isaacson, R. M., & Was, C. A. (2010a). Believing you're correct vs. knowing you're correct: A significant difference? *The Researcher, 23*(1), 1-12.

Isaacson, R. M., & Was, C. A. (2010b). An Educational Psychology Curriculum to Teach Metacognition. *The National Teaching and Learning Forum, 19*(5), 1-4.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*(3), 349-363. doi: 10.1037//0096-3445.131.3.349

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: the influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*(3), 303-314. doi: 10.1007/s11409-011-9083-7

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*(1), 7-28.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy and self-efficacy. *Metacognition and Learning, 1*(2), 159-179. doi: 10.1007/s10409-006-9595-6

Serra, M . J., & Metcalfe, J. (2009) Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education.* (pp. 278-298). New York, NY: Routledge.

Tobias, S., & Everson, H. (2002). *Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring*. College Board Report No. 2002-3. College Board, NY.

Tobias, S., & Everson, H. (2009). The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education*. (pp. 107-128). New York, NY: Routledge.

Vadhan, V., & Stander, P. (1994). Metacognitive ability and test performance among college students. *The Journal of Psychology, 128*(3), 307-309.

Was, C. A., Isaacson, R. M., Beziat, T.L.R., & Dippel, K. (2011, October). Practice Improves Absolute Accuracy of Knowledge Monitoring: Explicit Training to Overcome

Confidence Bias. Paper accepted for presentation at the *29th Annual Conference of the Northern Rocky Mountain Education Research Association.* Jackson, WY.

Correspondence regarding this manuscript should be sent to: Christopher A. Was, Kent State University, Educational Psychology, 405 White Hall, Kent, OH 44242. Phone: (330) 672-2929, Fax: (330) 672-2512, CWAS@KENT.EDU

Footnotes

1. A paired samples *t*-test indicated a significant difference between scores on Exam 1 and Exam 12, $t(206) = -6.25$, Mean Difference = 6.00, $p < .001$, CI = -7.89; -4.11.

Table 1. Means of Exam Scores, Predicted scores, and Calibration Across the Semester.

Exam	Mean Score	Mean Predicted	Calibration
1	76.52 (13.49)	82.66 (7.72)	5.45 (11.78)
2	82.56 (9.80)	84.62 (8.14)	2.06 (8.44)
3	80.26 (12.41)	83.22 (8.31)	2.92 (10.49)
4	80.80 (10.72)	83.67 (7.90)	3.00 (10.09)
5	79.34 (13.11)	83.79 (8.46)	4.12 (9.62)
6	87.04 (10.25)	85.34 (7.56)	-2.61 (8.71)
7	79.63 (10.31)	84.89 (8.62)	4.75 (9.27)
8	82.64 (13.71)	85.53 (8.00)	1.95 (9.93)
9	83.07 (12.05)	82.72 (8.51)	-.30 (9.79)
10	84.94 (13.24)	83.82 (8.72)	-1.42 (10.39)
11	84.47 (11.40)	84.83 (8.71)	.13 (9.01)
12	82.88 (11.97)	84.00 (8.31)	1.14 (9.82)

*Note: Standard deviations in parentheses.

Table 2. One Sample t-Tests of Calibration Mean of the Twelve Exams Compared to Zero.

Exam	Mean Calibration	*t*	df	*p*	95% CI
1	5.47	7.33	247	> .001	4.00; 6.94
2	2.17	3.84	235	> .001	1.06; 3.28
3	2.79	4.12	240	> .001	1.46; 4.12
4	3.36	4.83	229	> .001	1.99; 4.73
5	3.97	6.07	212	> .001	2.68; 5.26
6	-2.47	-4.10	203	> .001	-3.65; -1.28
7	4.65	7.13	210	> .001	3.37; 5.94
8	2.09	2.99	198	.003	.71; 3.47
9	-.39	-.59	217	.553	-1.66; .89
10	-1.56	-2.21	215	.028	-2.95; -.17
11	.04	.67	177	.947	-1.27; 1.36
12	1.12	1.76	210	.080	-.14; 2.37

Figure Captions.

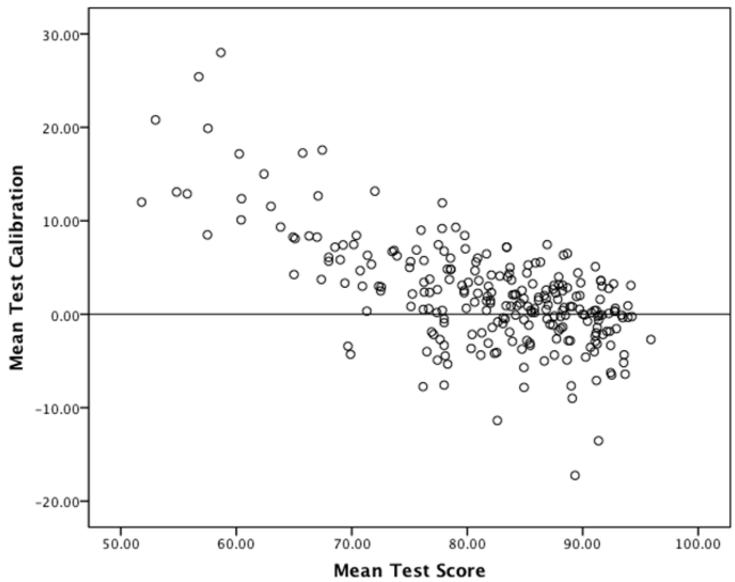

Figure 1. Mean test calibration by mean test score averaged across the semester.

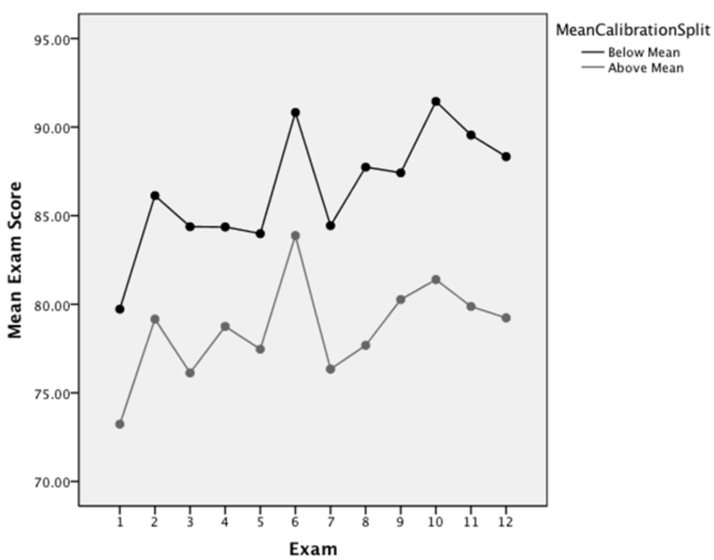

Figure 2. Mean exam scores across the semester by mean calibration split.

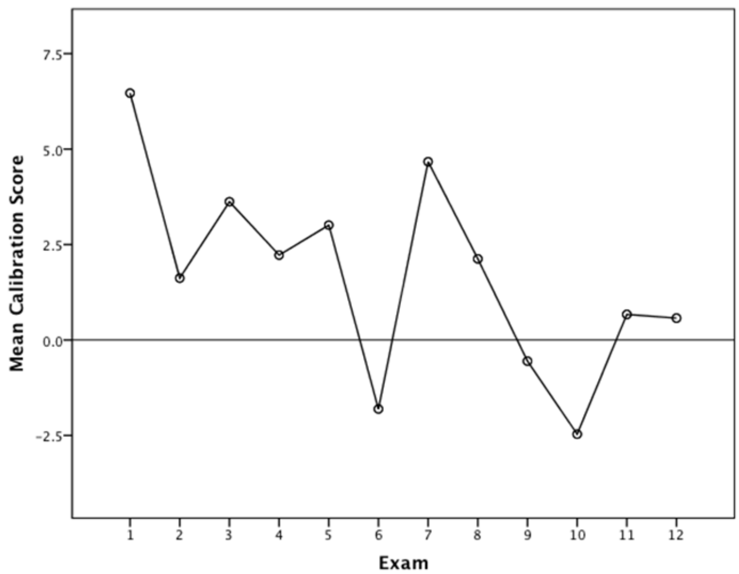

Figure 3. Mean calibration score by exam.

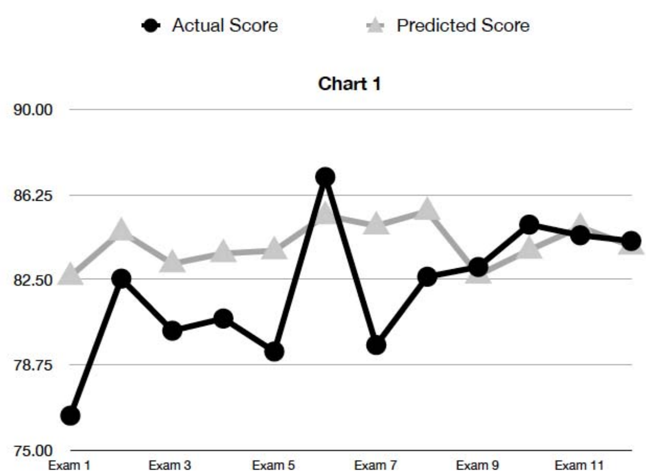

Figure 4. Mean predicted score and mean actual score by exam.